Fine Grained Categorization of Drug Usage Tweets

Priyanka Dey[1] and ChengXiang Zhai[1]

University of Illinois at Urbana-Champaign, Champaign IL, USA
pdey3@illinois.edu

Abstract. Drug misuse and overdose has plagued the United States over the past decades and has severely impacted several communities and families. Often, it is difficult for drug users to get the assistance they need and thus many usage cases remain undetected until it is too late. With the booming age of social media, many users often prefer to discuss their emotions through virtual environments where they can also meet others dealing with similar problems. The widespread use of social media sites creates interesting new opportunities to apply NLP techniques to analyze content and potentially help those drug users (e.g., early detection and intervention). To tap into such opportunities, we study categorization of tweets about drug usage into fine-grained categories. To facilitate the study of the proposed new problem, we create a new dataset and use this data to study the effectiveness of multiple representative categorization methods. We further analyze errors made by these methods and explore new features to improve them. We find that a new feature based on tweet tone is quite useful in improving classification scores. We further explore possible downstream applications based on this classification system and provide a set of preliminary findings.

Keywords: Categorization · Social media analytics · Drug usage · Public health

1 Introduction

The epidemic of drug misuse and overdose has left several communities and families devastated across the United States. Deaths from drug overdoses have risen sharply in recent years as shown in Figure 1. According to the Centers for Disease Control and Prevention (CDC), there have been almost 1 million deaths induced from drug overdose since 2000 [1]. Moreover, research from the CDC shows that drug usage is becoming increasingly detrimental; the 2020 rate of drug overdose deaths has accelerated and increased 31% since 2019 [1].

Although this issue is widespread, it is difficult to find ways to help such individuals. One of the biggest barriers in reducing this epidemic is that many users prefer to hide their drug addictions from others. Users may fear criminalization or perhaps fear judgment. Thus, it becomes increasingly difficult to identify patterns in drug usage and oftentimes many such drug usage cases remain undetected until it is too late. However, recent state-of-the-art research shows that

social media can help in identifying drug discussion. With the booming age of technology and social media, members of a community often prefer to discuss their emotions through a virtual environment (e.g. social media) where they can also meet others dealing with similar problems. [2], [3], [4] have found that users often discuss the abusive use of prescription drugs on social media sites such as Twitter.

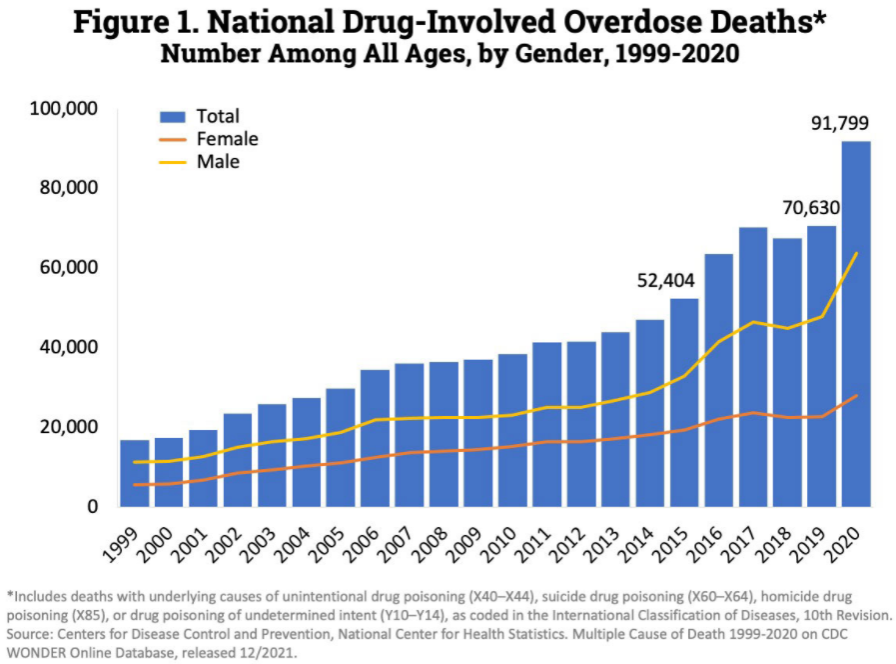

Fig. 1: Total Deaths by Overdoses from 1999-2020 in the US

Much of past research work has focused on identifying drug abuse on social media. One research work [2] showed that using n-grams and specified drug-related keywords (using a drug-slang lexicon, synonym expansion, and word clusters) allows for the creation of an effective system to classify tweets as "abuse" or "non-abuse" tweets. Another work [3] presented a supervised machine learning technique to identify prescription-drug abuse tweets. [5] presented a monitoring system of tweets using machine learning to identify drug abuse in real-time. Another work [6] focused on opioid abuse and designed a binary classifier for relevant tweets. Although such tweets facilitate the study of drug-abuse patterns, this coarse categorization does not allow us to fully understand and capture the myriad effects drug usage induces on a community. Thus, a more fine-grained categorization of tweets discussing drug usage is necessary. This further classi-

fication can then help us better study a community (e.g. crimes influenced by various drugs, reactions and opinions of users about drug usage).

To address this limitation, we present a novel task of identifying drug usage on social media and study how to categorize such tweets into fine-grained categories to enable many downstream applications such as real-time monitoring of drug users in a community, companion agents for drug users to help them on a daily basis, and drug-induced crime hotspot models. To facilitate this study, we generate a novel dataset consisting of drug usage tweets and study the effectiveness of multiple representative categorization methods using this dataset. We further analyze errors made by these methods and explore new features to improve them. We find that a new feature based on tweet tone is effective in improving classification scores. We then explore possible downstream applications based on this classification system and provide a set of preliminary findings.

2 Research Aim

The main contribution of this paper is to introduce and formulate a new tweet categorization problem so as to obtain a more detailed understanding of the tweets discussing drug usage. As previous work has studied how to identify tweets discussing drug abuse, we define our problem to further classify the tweets into refined categories. To study this, we introduce the following research questions:

1. How should we create a fine-grained categorization of drug-usage tweets?
2. How can we construct a new dataset to facilitate quantitative evaluation of classification techniques?
3. Which machine learning algorithms work best for this new task?
4. How can we leverage this categorization system to support further applications?

In the following sections, we first look at RQ1 and RQ2 where we detail how we have designed the various labels for drug usage and created our dataset. We then continue with RQ3 to identify the best algorithms to assist in the classification task. We follow with RQ4 and provide examples of possible use cases and some preliminary findings. We conclude with potential future research directions.

3 Designing Fine-Grained Labels for Drug-Usage Tweets

For RQ1, we first propose a new drug usage taxonomy which helps to better capture drug discussion in a multitude of aspects. After a manual analysis of drug-related tweets, we identified 3 broad categories to describe this discussion. Firstly, many tweets discuss the effects of drug usage in a community, i.e. drug-induced crimes or statistics related to drug usage in a community. We categorize such tweets as News/Research as they provide a recollection of events and statistics related to drug usage in a community. Tweets in this category

are particularly useful as they can give us insight into which areas of a state have common drug-induced crimes e.g. theft or battery. Authorities can then use these annotations to determine future crime hotspots and help reduce drug-related casualties.

From our manual analysis, we also observed that many tweeters discuss their personal and active use of drugs. Thus, we use Usage as our second category. Detection of tweets discussing drug usage provides several applications. For example, if a user makes a post about using a drug, health officials can monitor the user and prevent them from a possible overdose. Furthermore, psychologists can learn more about a patient's behavior based on the user's tweets posted while under the influence. In order to further understand usage, we divide this category based on a temporal aspect i.e. Past Usage, Current Usage, and Intent to Use. This further subcategorization of usage tweets enables more effective monitoring. For example, health and addiction officials can build a real-time monitoring system to determine if a tweet discusses current usage. Officials can then extract tweet information such as the user's region and determine the next best steps in order to prevent a drug overdose and/or casualty. Past usage tweets can provide cues as to whether a person plans on using again; Intent to Use tweets can help officials take measures to provide early, preventative care.

We also identified a third category "Reactions/Opinions" based on tweets that express opinions and reactions regarding drug usage. Identification of such tweets can be useful, e.g. psychologists can monitor annotated tweets classified as reactions/opinions to see how the members of a community feel about some event or general drug use (e.g. members conversing about the legalization of marijuana in a state). These tweets can then provide insight into members' possible future actions (e.g. more drug usage). Tools designed to help these individuals (e.g. companion chatbots) can then be built based on these data.

To summarize, we present a total of 6 fine-grained categories:

1. **News/Research**: Discuss any events related to drug use (e.g. arrests, crimes)
2. **Usage**: Discuss user using drugs. These tweets are further sub-classified into 3 categories:
   (a) **Past Usage**: Refer to events where the user has used in the past
   (b) **Current Usage**: Refer to user using (at the time the tweet was written)
   (c) **Intent to Use**: Refer to those who are planning on using in the near future
3. **Reactions/Opinions**: Refer to a response (reaction or opinion) to some drug-induced act

In order to generate a classification system for these categories, we design two separate classification models: a topical model to understand tweet context (News/Research, Usage, Reactions/Opinions) and a temporal model to further understand drug usage (Past Usage, Current Usage, Intent to Use).

4 Dataset Construction

This research is a first attempt to study fine-grained drug usage on social media; thus, for this task, we generate our own dataset. We collect drug usage tweets using the Twitter Search API. We gather tweets from November 2016 to December 2016; we filter based on drug keywords which include names of drugs e.g. heroin, marijuana, weed, cocaine, fentanyl, ketamine, methamphetamine and words related to addiction and abuse such as opioidcrisis, addiction, overdose, compulsions.

We identified approximately 5% of the collected tweets between the 2 months to be relevant for our study. Table 1 shows the 10 most frequently found drug-related words in the tweets. After collection, all tweets were manually annotated into one of the categories listed in the previous section. Table 2 shows the number of tweets per category and corresponding example tweets. In total, we have generated a dataset of 921 annotated tweets.

Table 1: Most Common Words in Drug-Related Tweets

Word	**Frequency**	**Relative Frequency**
Weed	519	38.73%
Marijuana	238	17.76%
Addiction	161	12.01%
Cannabis	80	5.97%
Cocaine	78	5.82%
Opioid	76	5.67%
Heroin	68	5.07%
Booze	53	3.96%
Overdose	43	3.21%
Nicotine	11	0.82%

5 Classification Methods

As an initial study of the new task and for RQ3, we set our goal as to establish some baseline results by evaluating a few popular representative multi-class classifiers that have shown success in several classification tasks.

1. Multinomial Naive Bayes (NB): modification of the Naive Bayes model and is popularly used for text document classification

2. Logistic Regression (LogR): uses the logistic function to determine whether or not data falls into a particular category

3. Support Vector Machine (SVM): goal is to calculate the maximal margin hyperplane(s) separating the categories of the data

4. Random Forests (RF): uses an average of decision tree predictions to determine the best predictive accuracy

Table 2: Tweet Categories: Examples and Counts

Tweet Category	**Example**	**Count**
News	"Ex-Ravens cheerleader charged with rape, supplying booze to minor set for hearing"" Remember, rape isnt always male to female."	301
Past Usage	"I told my mom I do methamphetamine and she responded with ""So thats why your face is so f**ked up"	104
Current Usage	"smokin weed in the backstage photo booth with Duchovny."	130
Intent to Use	"Ill find some weed tonight."	195
Reactions	"in debate we had to choose from 4 topics and i chose marijuana. Im debating on the CON side because marijuana is SO stupid."	191

5. Multi-Layered Perceptron (MLP): feed-forward artificial neural network where training data is split into multiple layers of information.
6. Extremely Randomized Trees (Extra): conceptually similar to Random Forests but splitting the tree is based on a random attribute rather than highest information gain.
7. Bidirectional Encoder Representations from Transformers (BERT): transformer-based machine learning model popularly used in NLP tasks and can be fine-tuned for various downstream tasks. Research has shown this model can work quite well for various classification tasks including text using BERT contextualized embeddings [7].

We generated features using word2vec [8] by transforming tweets into a set of vectors and then computing a mean vector for each tweet for the first 6 models. We used the default BERT embeddings for the BERT model. We then used the Python sklearn package [9] to train each classifier to best fit the training data. 80% of the annotated dataset was used as training data, and the remaining 20% was used for testing. 5-fold cross-validation was used to calculate the accuracy, precision, and recall scores.

5.1 Comparison of Classifiers for Topical Model

The results for the topical model are summarized in Table 3. The News and Usage categories were the easiest to categorize as can be seen by their relatively high values of accuracy, precision, and recall scores. The SVM seemed to work best for the News Category whereas the LogR and NB models worked best for the Usage category. The Reactions category, however, proved to be difficult to classify with most classifiers having extremely low recall scores. The best models appear to be BERT, the RF, and the Extra. Overall, the BERT model performed the best across all categories and provided the highest scores for the Reactions category.

Table 3: Class Accuracy, Precision, and Recall Scores for Topical Model

Classifier	News			Usage			Reactions		
	Acc	Precision	Recall	Acc	Precision	Recall	Acc	Precision	Recall
Naive Bayes	0.65	0.94	0.80	**0.77**	**0.80**	**0.91**	0.39	0.50	0.16
Random Forest	0.68	0.70	0.86	0.64	0.64	0.94	0.48	1.00	0.32
SVM	**0.81**	**0.88**	**0.86**	0.65	0.64	0.94	0.41	0.64	0.17
LogR	0.71	0.70	0.92	**0.71**	**0.76**	**0.86**	0.30	0.40	0.19
MLP	0.77	0.77	0.92	0.69	0.68	0.95	0	0	0
Extra	0.68	0.66	0.94	0.68	0.70	0.86	0.45	0.72	0.34
BERT	0.79	0.79	**0.88**	**0.90**	**0.90**	0.81	**0.54**	**0.54**	**0.61**

5.2 Comparison of Classifiers for Temporal Model

Results for the temporal model are summarized in Table 4. This classification proved to be a much more difficult task. Intent to Use and Current Usage were easiest to identify as seen by their relatively higher accuracy, precision, and recall scores. The RF and Extra models seem to work best for these categories. The LogR model also worked well for classifying Intent to Use tweets. On the other hand, the Past Usage was extremely difficult to identify, and most classifiers did not work well, suggesting that there is much room for future research on this categorization task. Overall, the LogR model provided high scores for all categories.

Table 4: Class Accuracy, Precision, and Recall Scores for Temporal Model

Classifier	Past			Current			Possible Intent to Use		
	Acc	Precision	Recall	Acc	Precision	Recall	Acc	Precision	Recall
Naive Bayes	0.29	0.10	0.17	0.51	0.40	0.71	0.53	0.79	0.52
Random Forest	0.25	0.10	0.12	**0.52**	**0.41**	**0.70**	0.54	0.78	0.52
SVM	0.40	0.24	0.42	0.44	0.32	0.57	0.56	0.84	0.54
LogR	0.42	0.19	0.37	0.48	0.40	0.59	**0.69**	**0.74**	**0.80**
MLP	0.31	0.10	0.20	0.48	0.24	0.75	0.51	0.84	0.48
Extra	0.33	0.13	0.33	**0.53**	**0.40**	**0.77**	0.53	0.84	0.50
BERT	0.21	0.15	0.18	0.42	0.47	0.44	0.41	0.29	0.34

5.3 Error Analysis & Novel Tweet Tone Feature

After determining the best ML algorithms for the topical and temporal models, we explored examples of misclassified tweets to obtain insights for possible solutions to improve classification performance.

For each of the categories, we have identified some common tweets misclassified by many of the ML algorithms. We present some examples in Table 5. Based

on a manual lexical analysis of these tweets, we found that each of the categories seemed to have a unique emotion or tone exhibited by their tweets.

For example, News/Research tweets tend to discuss drug-induced crimes or difficult stories about recovering addicts, both of which are despairing topics. On the other hand, emotions such as joy can be seen in Usage (e.g. a tweeter was able to buy his favorite weed); a pensive tone can be seen in Reactions/Opinions tweets (e.g. lack of support for legalizing weed). Based on these findings, we explored the use of a novel tweet categorization feature - tweeter tone (detected by a user's choice of words in a tweet).

Table 5: Examples of Misclassified Tweets

Category	**Tweet**
News	"kilo of cocaine only worth 1500 in Columbia $77k in Britain tho"
Past Usage	"i smell like pussy money n weed n she said Ooo i like yo cologne"
Current Usage	"I think somehow we all took a syph of Toms weed and are just tripping the **** out"
Intent to Use	"lmfao this sophomore turned to me and asked for legal advice about smoking and i just gave him so bs answers"
Reactions	"if kids these days understood the effects drugs can have on your brain, maybe the world would be a better place"

For each category, each training tweet's tone was determined using the Watson Tone Analyzer API [10] which categorizes a tweet's tone as anger, fear, joy, tentative, analytical, sad, confident, or none. This additional tone feature was then added to each training and testing sample's feature vector. The ML algorithms were then retrained using the updated feature vector to measure the new classification scores. We specifically chose to analyze the BERT and LogR models since they provided the best scores for the topical and temporal models. Tables 6 and 7 show a comparison of the precision and recall scores with the addition of the tone feature.

From these tables, we can see that the tone feature can indeed lead to performance improvement in both precision and recall. The News/Research category has an extremely high recall as well as a great improvement in precision. The Reactions/Opinions category has a higher precision score as well. Most usage categories are still difficult to identify, but the tone feature did not seem to lower the precision and recall scores significantly. Thus based on this experiment, it seems the tone of a tweet message can improve scores for this classification task.

6 Exploration of Possible Applications Using Fine-Grained Drug Usage Tweet Classification

The classification models and fine-grained drug usage taxonomy that we have introduced can further lead to the creation of downstream applications. Although

Table 6: Comparison of Precision and Recall Scores in Topical Model (BERT) w/ Tone Feature

Category	Original Precision	New Precision	Original Recall	New Recall
News	0.79	0.88	0.85	0.91
Usage	0.90	0.90	0.81	0.83
Reactions	0.54	0.74	0.61	0.74

Table 7: Comparison of Precision and Recall Scores in Temporal Model (LogR) w/ Tone Feature

Category	Original Precision	New Precision	Original Recall	New Recall
Past Usage	0.19	0.33	0.37	0.43
Current Usage	0.40	0.42	0.41	0.52
Possible Intent	0.74	0.87	0.80	0.78

a full exploration of possible applications is beyond the scope of this paper, we propose 3 different future applications and perform some preliminary work for each:

1. **Lexical Analysis of Usage Tweets**: Illicit drug usage is oftentimes difficult to identify. By viewing Usage tweets containing illicit drug keywords, we may be able to find related drug cases.
2. **Sentiment Analysis of Reactions/Opinions Tweets**: Such an analysis can be beneficial in predicting community changes: legalizing marijuana for recreational use.
3. **Geographical Analysis of Usage Tweets**: This analysis can be useful in determining whether certain areas of the United States have similar drug usage patterns.

In order to perform initial experiments for each of the proposed applications, we first use our topical model (BERT) and temporal model (LogR) to classify additional tweets. We classify an additional 14,746 tweets gathered from the Twitter API between the months January to March of 2021. We were able to classify a total of 4,850 News/Research tweets, 8,506 Usage tweets (3,534 Past Usage, 3,180 Current Usage, and 1,92 Intent to Use), and 1,390 Reactions/Opinions tweets. We then extract locations from these tweets based on users' locations using the Twitter API. We were able to obtain 13,189 geo-tagged tweets which we use for the proposed applications' analyses.

6.1 Lexical Analysis of Usage Tweets

One possible application is analyzing the classified usage tweets to determine the usage of illicit drugs. Illicit drug usage can oftentimes be difficult to identify as users may fear imprisonment or other types of criminalization. However, as shown through Sarker et al. and Shutler et al., many users do indeed discuss

illegal drug usage on social media. Thus, further analysis of Usage tweets can help identify patterns. Moreover, the fine-grained analysis of Usage tweets can give us an additional temporal feature.

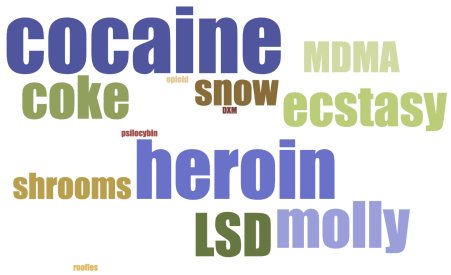

Fig. 2: Word Cloud for Most Common Illicit Drugs in Drug Usage Tweets

To identify illicit drug usage in the United States via the Usage tweets, we first filter tweets that contain names and slang for illegal drugs (e.g. "coke", "roofies", "cocaine"). In Figure 2, we present a word cloud of the most commonly discussed illegal drugs. We identify "heroin" and "cocaine" to be the most commonly misused drugs. We then generate a choropleth map to further show which states have the most illicit drug usage tweets. We present our findings in Figure 3. Based on this map, we see that Idaho, Indiana, and Maine have the highest illicit drug usage tweets in the United States.

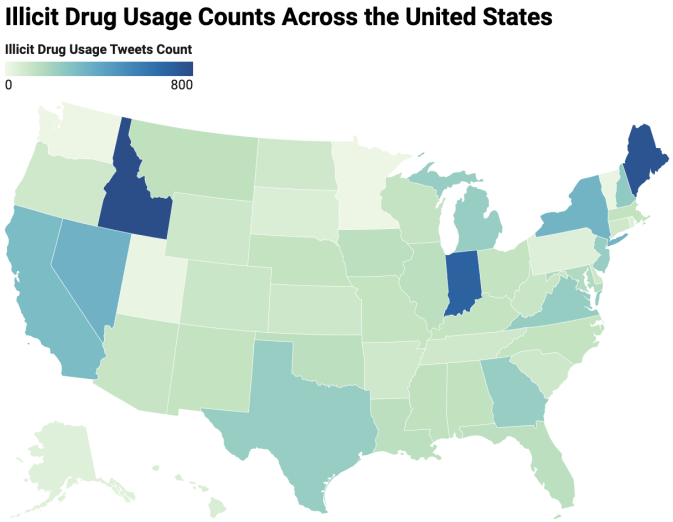

Fig. 3: Choropleth Map of United States: Counts of Illicit Drug Usage Tweets

6.2 Sentiment Analysis of Reactions/Opinions Tweets

Another possible application is to analyze the classified Reactions/Opinions tweets to determine the public's opinion on drug usage. Such an application can be particularly useful when deciding effective drug control policies. For example, if marijuana usage is favorable in a particular region, severe legal restrictions against weed might instigate an increase in its black market.

To determine sentiments, we leverage the VADER (Valence Aware Dictionary for Sentiment Reasoning) python package [11] which provides scores on a document for various sentiments. We apply this method to the 13,189 geo-tagged tweets we obtained from further classification and extract scores for the positive and negative sentiments. To calculate sentiment for a US state, we compute the mean of sentiments for all tweets retrieved from a state. We present our findings in a choropleth map in Figure 4; positive scores show an overall positive reaction towards drug usage in that state and negative scores show an overall negative reaction. Some states also had mixed opinions towards drug usage and thus mean scores around 0. We identify Indiana, Virginia, and Nebraska to have the highest positive sentiments and Wisconsin, Arkansas, and Alabama to have the highest negative sentiments.

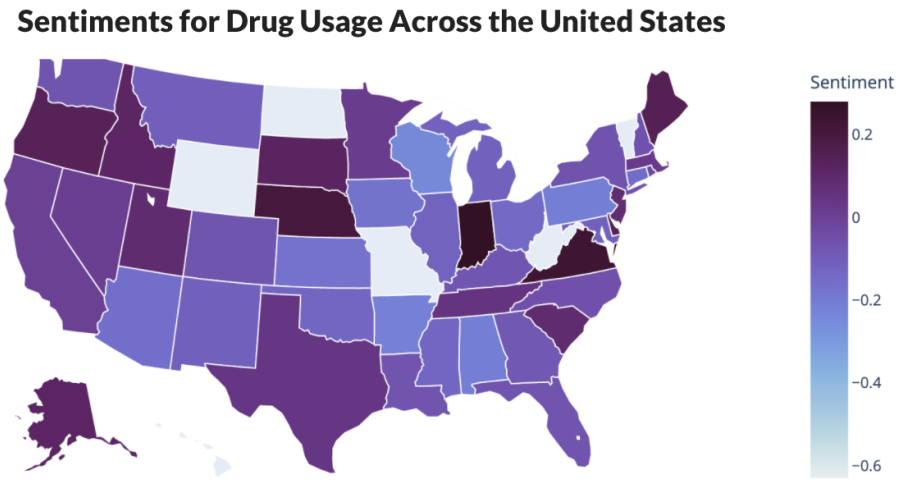

Fig. 4: Choropleth Map of United States: Sentiments of Drug Usage Extracted from Reactions/Opinions Tweets

6.3 Geographical Analysis of Usage Tweets

A third possible application is to understand and determine drug usage patterns across the United States. Oftentimes, neighboring states may have similar drug

usage rates. We seek to determine location-based patterns on drug usage. To achieve this, we again use the 13,189 geo-tagged tweets and observe drug usage tweet counts for each state. We present results in Figure 5. Although many states have similar counts (with most tweets centered around California, Texas, Florida, and New York), we can see that many of the Southeastern states and Mid-western states have similar amounts of Usage tweets.

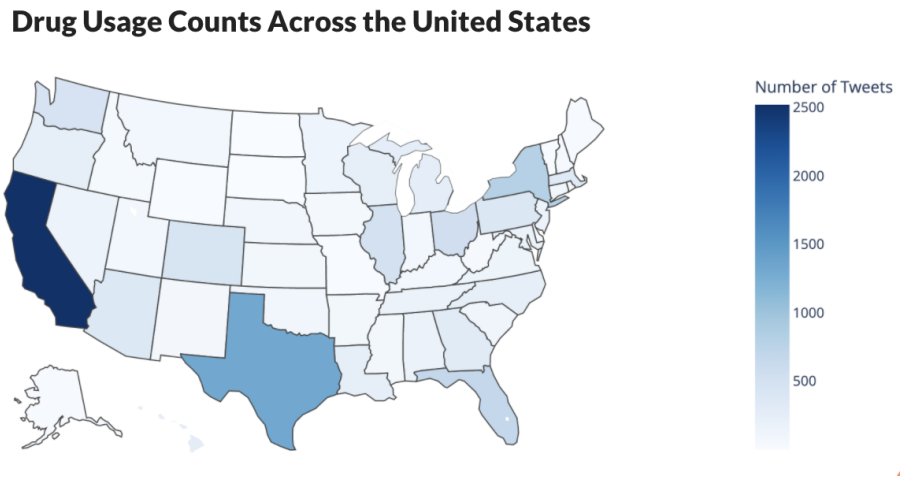

Fig. 5: Choropleth Map of United States: Drug Usage Counts Across the United States

Although not currently explored in this preliminary analysis, the fine-grained categorization of Usage tweets further allows us to monitor real-time patterns for Current Usage and Intent to Use tweets.

7 Conclusion

In this paper, we introduced a novel and useful classification task to better understand drug discussion on Twitter by categorizing tweets into fine-grained categories: News/Research, Usage (Past Usage, Current Usage, and Intent to Use) and Reactions/Opinions. We have created a new annotated dataset which we use to evaluate several popular representative categorization methods (NB, LogR, SVM, RF, MLP, Extra, BERT) . We further analyzed common errors and identified tweet tone to improve classification results. We then explored some downstream applications and presented some initial work based on our proposed classification system.

One limitation of our work is that the dataset we have constructed is small, which is mainly due to the limited resources available to us. Although we are able to make some interesting findings in this paper, an important future work

is to further construct larger datasets. Our baseline results also show that some categories, notably Past Usage, are difficult to classify, and in general, there is much room for additional research to improve the accuracy of this task.

Finally, the trained classifiers using our annotated data set can already be potentially useful for building novel applications as we have shown in Section 6. This work can further be extended to generate a real-time system for monitoring drug usage tweets. Such a system can then be used to extract common emotions among users. These tones can then be used to build a companion chatbot that can provide users with a customized experience. These common emotions can also be used to build a system that allows psychologists to monitor behavior patterns in drug users and in turn help them develop effective and personalized drug treatments.

8 Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1801652 and by the National Institutes of Health under Grant 1 R56 AI114501-01A1.

References

1. CDC: "Now Is The Time To Stop Drug Overdose Deaths" Article, https://www.cdc.gov/drugoverdose/featured-topics/overdose-prevention-campaigns.htm. Last accessed 1 Feb 2022
2. Sarker, A., O'Connor: Social Media Mining for Toxicovigilance: Automatic Monitoring of Prescription Medication Abuse from Twitter. Drug Saf 39, 231–240 (2016). https://doi.org/10.1007/s40264-015-0379-4
3. Shutler, L., Nelson, L.: Drug Use in the Twittersphere: A Qualitative Contextual Analysis of Tweets About Prescription Drugs. J Addict Dis. 34(4), 303-310 (2015)
4. Chary, M., Genes, N.: Leveraging Social Networks for Toxicovigilance. J Med Toxicol. 9(2), 184-191 (2013).
5. Phan, N., Chun, S.: Enabling Real-Time Drug Abuse Detection in Tweets. In: 2017 IEEE International Conference on Data Engineering (ICDE). pp. 1510-1514, https://doi.org/10.1109/ICDE.2017.221
6. Flores, L., Young, S.: Regional Variation in Discussion of Opioids on Social Media. Journal of Addictive Diseases. 39(3), 315-321 (2021). https://doi.org/10.1080/10550887.2021.1874804
7. Sun, C., Qiu, X.: How to Fine-Tune BERT for Text Classification? Chine Computational Linguistics, 194-206 (2019).
8. Mikolov, T., Chen, K.: Efficient Estimation of Word Representations in Vector Space. In: Bengio, Y., LeCun, Y. (eds.) 1st International Conference on Learning Representations (ICLR) 2013, Workshop Track Proceedings, IEEE, Arizona, USA.
9. Pedregosa, F., Varoquaux, G.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. (12), 2825-2830 (2011).
10. Agrawal, A., Sonawane, S.: Tone Analyzer. International Journal of Engineering Science and Computing. 7(10), 15060-15064 (2017).

11. Hutto, C., Gilbert, E.: VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In: Proceedings of the International AAAI Conference on Web and Social Media, ICWSM, vol. 8, pp. 216-225, Michigan, USA. (2014)